THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 3, March 1982
Printed in U.S.A.

Note on Some Factors Affecting Performance of Dynamic Time Warping Algorithms for Isolated Word Recognition

By L. R. RABINER

(Manuscript received October 14, 1981)

To implement a dynamic time warping (DTW) algorithm for isolated word recognition, a number of factors must be specified. These factors include endpoint constraints, local path constraints, global path constraints, axis orientation, and local distance measure. Although a number of studies have been made to decide how best to choose these factors, several unresolved issues remain. The problems with choosing these factors become more complicated if the word reference patterns are not created from whole words, but instead are built up from subword units, e.g., syllables or demisyllables. In this paper, we consider an isolated word recognizer in which the reference patterns are obtained from a set of demisyllables as specified in a user-supplied lexicon. Different pronunciations of a word are represented by multiple entries in the lexicon. Because of the inherent boundaries in the reference patterns (at each demisyllable junction), it was felt that local constraints could influence performance more than for the standard whole-word case. Further, it was felt that some flexibility in the endpoint path constraint at the end of the word would be helpful, in general, for isolated word recognition caused by high variability at the ending of words with stops. A DTW algorithm was programmed and tested on an 1109-word vocabulary. Results showed small accuracy improvements when the path endpoint was allowed to vary across four frames (of either test or reference pattern). Loosening the local path constraints, however, had a significant degrading effect on the performance.

I. INTRODUCTION

The technique of using dynamic time warping (DTW) to align (in time) a test and reference pattern for isolated word recognition has

been shown to be effective in a wide variety of recognition systems.[1-6] Although a great deal of investigation has been made into "optimal" DTW algorithms, there still remains uncertainty as to how best to specify the factors of the DTW implementation to achieve the highest recognition accuracy.

In this paper, we consider two of these factors, namely, the endpoint constraint at the end of the warp and the local path constraints. Previous investigations were made by Rabiner et al.[4] and Myers et al.[6] on the effects on word recognition accuracy of both loosening endpoint constraints and using different local path constraints. However, the work[4] on DTW algorithms with relaxed endpoint constraints considered only two specific variations, namely, the unconstrained endpoint case (the UE2-1 algorithm), and the local minimum case (the UELM algorithm). Neither of these algorithms considered relaxing just the endpoint constraint at the end of the word. This constraint is very important since replications of isolated words generally have the most variability at the end of the word. Similarly, although various local path constraints were considered in Ref. 6, the use of subword units in the reference patterns raises again the question as to whether increased flexibility in the choice of warping path leads to improvements in recognition scores. We show that by loosening the endpoint constraint at the end of the utterance, a small but consistent increase in recognition accuracy is obtained. It is also shown that the Itakura local path constraints yield higher recognition accuracies than generalized Itakura local constraints (Type III, Ref. 6). This result corroborates previous findings[4,6] which show that, in general, opening up of the region for DTW matching of words generally leads to degraded performance, since the improved score for the correct matches is offset by the improved scores for incorrect matches.

The outline of this paper is as follows. In Section II, we briefly review the DTW implementation, and describe the factors whicb were studied. In Section III, we present and discuss results of a recognition test using a 1109-word vocabulary, and in Section IV, we summarize the findings.

II. THE DTW IMPLEMENTATION

Assume that we are given a test pattern T, consisting of a sequence of N vectors, i.e.,

$$\mathrm{T} = \{T(1), T(2), \cdots, T(N)\}, \tag{1}$$

where the vector $T(i)$ is a spectral representation of the ith frame of the test word. The vectors, $T(i)$, are a set of 9 autocorrelations (from which an 8th order linear predictive coding model is derived). The

duration of the test word is N frames, where each frame represents 45 ms of speech, and adjacent frames are spaced 15 ms apart.

For a given vocabulary of V words, we denote the reference pattern for the vth word as $\mathbf{R}_v$, and we represent each reference pattern as a sequence of M_v vectors, i.e.,

$$\mathbf{R}_v = \{R_v(1), R_v(2), \cdots, R_v(M_v)\}, \tag{2}$$

where each vector is again a spectral representation of the corresponding frame within the word.

To optimally align the time scales of the test (the n index) and reference (the m index) patterns via a DTW algorithm, we must solve for a warping, or path alignment function of the form

$$m = w(n) \tag{3}$$

and thereby seek to minimize the average distance

$$D = \frac{1}{N} \sum_{n=1}^{N} \tilde{d}\{T(n), R[w(n)]\} \tag{4}$$

over all possible $w(n)$, where $\tilde{d}$ is the local distance between test frame n and reference frame $m = w(n)$. To solve the DTW problem requires specification of the following:[6]

(*i*) endpoint constraints
(*ii*) local path constraints
(*iii*) global path constraints
(*iv*) axis orientation
(*v*) local distance measure.

In this paper, we have considered the effects, on recognition accuracy, of two of the DTW factors—the local endpoint constraints, and the local path constraints. In particular, we have considered the following DTW specifications:

Endpoint Constraints—We assume the word endpoints of the test and reference patterns satisfy the path constraints

$$w(1) = 1 \tag{5a}$$

$$M - \delta_R \leq w(n) \leq M, \qquad N - \delta_T \leq n \leq N, \tag{5b}$$

where δ_R is a range of frames, at the end of the reference pattern, and δ_T is a range for frames, at the end of the test pattern, where the warp path can terminate. Thus, the warp path begins at the point (1, 1) and ends in a region of δ_R frames from the end of the reference, and δ_T frames from the end of the test.

Local Path Constraints—We assume that the local path satisfies one of the two sets of constraints shown in Fig. 1, namely, the Itakura[2]

path constraints (Fig. 1a) and the Myers et al.[6] Type III path constraints (Fig. 1b). Both of these local path constraints satisfy the continuity equations

$$0 \le w(n) - w(n-1) \le 2 \tag{6a}$$

$$w(n) - w(n-1) = 0 \Rightarrow w(n-1) - w(n-2) > 0. \tag{6b}$$

The difference between the Itakura and the Type III path constraints is a subtle, but an important one, and it is related to the look-ahead (or look-back) capability of the local path constraints in finding the best path to a given grid point. This difference is illustrated in Fig. 2 which shows three possible local paths to grid point (n, m). Path 1 comes initially from grid point $(n-2, m-1)$ and goes through grid point $(n-1, m)$ before terminating at grid point (n, m). Path 2 goes from grid point $(n-1, m-1)$ to grid point (n, m). Path 3 goes from grid point $(n-1, m-2)$ to grid point (n, m). For the Itakura constraints (with 1 level of look-ahead logic), if the best path to the intermediate grid point $(n-1, m)$ came from grid point $(n-2, m)$—that is, it came across—then the possibility of path 1 being the best path to grid point (n, m) is not considered, whereas for Type III constraints (with two levels of look-ahead logic) path 1 is considered along with paths 2 and 3. At the end of a path (and often within the path), these differences can be significant.

Global Path Constraints—The endpoint and local path equations constrain the range of frames of the reference pattern (the m-axis) for which the basic DTW recursion is done to:

$$M_L(n) \le m \le M_H(n), \tag{7}$$

where

$$M_H(n) = \min\{2(n-1) + 1, M - (N - \delta_T - n)/2, M\} \tag{8a}$$

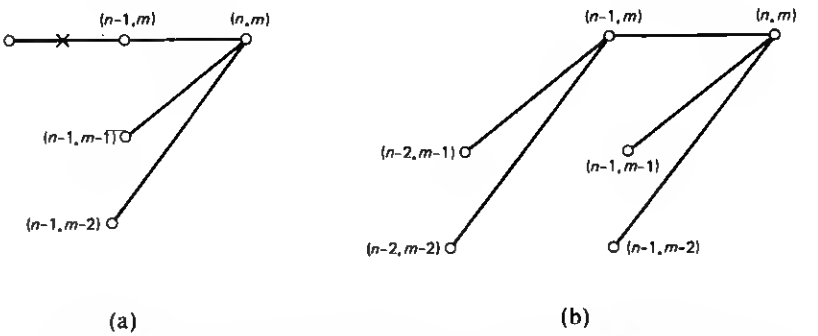

Fig. 1—Two sets of DTW local path constraints. (a) Itakura constraints. (b) Type III constraints.

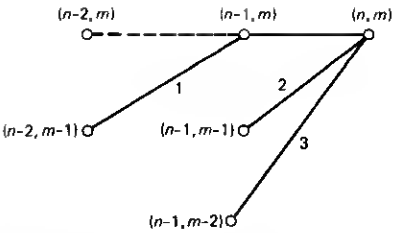

Fig. 2—Possible local paths for Itakura and for Type III local constraints.

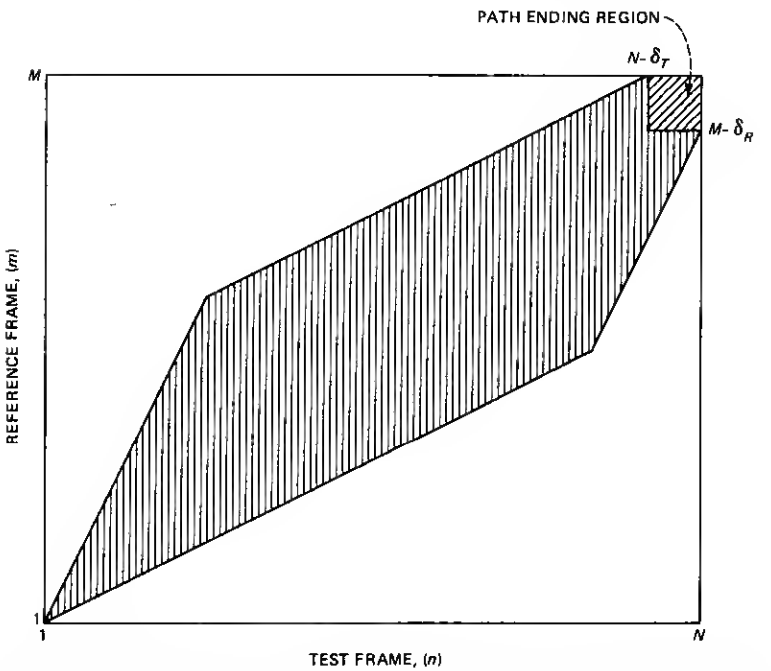

Fig. 3—The region in the warping plane in which the dynamic time warping path can lie with ending uncertainty of δ_T frames along the test and δ_R frames along the reference.

$$M_L(n) = \max\{(n-1)/2 + 1, M - 2(N-n) - \delta_R, 1\}. \quad (8b)$$

These global path constraints essentially define a parallelogram (as shown in Fig. 3) with lines of slope 2 and ½ emanating from the points $m = 1$, $n = 1$, and from $m = M - \delta_R$, $n = N - \delta_T$.

Axis Orientation—We assume that the test sequence index n is always mapped to the abscissa and the reference sequence index m is always mapped to the ordinate. Experience has shown that this orientation leads to the best performance in isolated word recognition systems.[4,6]

Local Distance Measure—The local distance measure used in this study is the Itakura log likelihood ratio[2] which is implemented in the form

$$\tilde{d}[T(n), R(m)] = \log[T(n) \cdot R(m)], \tag{9}$$

i.e., a log of the dot product of the two vectors $T(n)$ and $R(m)$.

2.1 Word recognition algorithm

Word recognition is achieved by computing the optimal warping path and distance for each word in the vocabulary, giving

$$D_v = \min_{w(n)} \min_{N-\delta T \le \hat{N} \le N} \left[\frac{1}{\hat{N}} \sum_{n=1}^{\hat{N}} \tilde{d}\{T(n), R_v[w(n)]\} \right] \tag{10}$$

and using the nearest neighbor rule to choose v^* as the best candidate where

$$v^* = \underset{v}{\operatorname{argmin}}[D_v]. \tag{11}$$

An ordered set of word distances is also maintained for statistical analysis purposes.

2.2 Creation of reference patterns

The reference patterns, R_v, for the vth vocabulary word was created from a lexicon which contained one or more demisyllable-based specifications for each word.[7] The entries in the lexicon were obtained by converting a standard dictionary word pronunciation to a set of demisyllable tokens which could be concatenated to create the word. Variant word pronunciations were represented by multiple lexical entries.

A set of 955-demisyllable tokens were used as the set of basis units from which each word was created. An 1109-word vocabulary (Basic English[8]) was used as the recognition vocabulary and a total of 1773 lexical entries were used for the vocabulary.

In creating the individual reference patterns from the demisyllable prototypes, a distinct boundary is created between each pair of demisyllables. To minimize boundary effects, a nonlinear smoother was used to interpolate (in a minimum mean-squared error sense) the vocal tract log area ratios (i.e., the ratio between the areas of adjacent sections of a nine-section vocal tract model for each frame of speech) in a four-frame vicinity of the boundary. Because of the presence of one or more boundaries within each reference pattern, it was felt that the local path constraints could potentially have more effect on the recognition results than for the case of reference patterns created from isolated words.

2.3 Summary of DTW factors that were studied

As discussed in this section, there are two DTW factors that were studied in the context of isolated word recognition from reference patterns created out of a corpus of demisyllables. These factors were as follows:

(*i*) *The range at the end of the test pattern*, δ_T—Two values of δ_T were chosen, namely, $\delta_T = 0$ (the standard case) and $\delta_T = 4$ frames (60 ms), corresponding to the duration of a release of a burst, etc.

(*ii*) *The range at the end of the reference pattern*, δ_R—For reasons similar to those in the choice of values for δ_T, values of 0 and 4 were studied for δ_R.

(*iii*) *The local path constraint*—Both the Itakura and the Type III constraints were studied in conjunction with variations of δ_R and δ_T. Only one simple experimental evaluation was made, and the results are given in Section III.

III. EXPERIMENTAL EVALUATION

To study the effects of the two factors discussed at the end of Section III on the recognition accuracy, an experiment was performed in the following way. The system was run as a speaker-trained, isolated word recognizer with a testing vocabulary of 1109 words (i.e., one replication of each word of the vocabulary) and a reference set of 1773 patterns obtained from the lexical description of the 1109 word-vocabulary. The DTW algorithm (with the flexibility of varying the DTW factors) was programmed in FORTRAN on a Data General Eclipse S230 computer. (Normally, a CSP MAP 200 array processor is used to process the data at high speed; however, the flexible version of the DTW algorithm was not available on the MAP.) Each DTW alignment took on the order of 1 second on the Eclipse; hence, a complete test on a *single* set of DTW factors would have taken on the order of 5000 hours of computing. Clearly, this amount of computation is prohibitive. Thus, an alternative testing procedure was derived in which a "standard" DTW was run on the CSP MAP (where $\delta_R = \delta_T = 0$ and the Itakura local constraints were applied). The complete matrix of distances (1109 words × 1773 references) was scanned and for each test word, the most highly confused alternative word was found (i.e., the word reference different from the test token with the smallest distance). A second "word recognition" test was now performed on the Eclipse in which, for each test word, only two DTW distances (with the variable factors included) were measured. The DTW distances were, thus, obtained for a reference corresponding to the spoken word, and for a reference corresponding to the vocabulary word most similar to the spoken word. In this manner, the eight combinations of δ_R, δ_T, and

local constraints were varied and their effects on word recognition accuracy were measured.

Before proceeding on to the results, some justification of this nonstandard testing methodology must be given. It should be clear that for test words which were correctly recognized, and which had one or fewer close alternative words, such a procedure is acceptable for studying the effects of DTW parameters. It can also be seen that for cases in which the DTW factors degrade performance using the two candidate algorithms, performance with all candidates will be degraded even further. Hence, the only cases in which we cannot place total reliance on the results are those words with several close candidates for recognition. Hopefully, the number of such cases is small and will not greatly affect the results presented.

3.1 Experimental results

For each set of DTW factors (i.e., specification of δ_R, δ_T, and LC—local constraints) a set of three quantities were measured, namely:

(*i*) *Overall recognition accuracy*—This quantity assumes that use of the two references best matching the test word was adequate when the DTW factors were varied. To the extent that this is the case, this measurement is the performance index of most concern.

(*ii*) *Average distance separation*—This quantity measures the average difference of the distances from the test to the reference pattern of the spoken word, and to the reference pattern of the closest competitor. It should be noted that the average distance separation was computed over *all* vocabulary words (including errors); hence, for errors the distance separation is negative and for correct recognitions it is positive. The average distance separation is one measure of separation of the distances for the correct words, and distances for nearest competitors.

(*iii*) *Number of occurrences of maximum distance separation over the eight factors*—This quantity counted the number of times the particular set of DTW factors provided the best distance separation (over the eight sets of factors) for a given word, if it was recognized correctly. Hence, this measurement is a marginal count (conditioned on correct recognition) of the superiority (in terms of distance separation) of one set of factors over the seven alternative sets.

The results of the word recognition experiment, in terms of the above three measurements, are given in Table I. The code $LC = 0$ is used to denote the Itakura local constraints; similarly, $LC = 1$ is used to denote the Type III local constraints. The following results can be seen from Table I:

(*i*) The overall recognition accuracy using $LC = 0$ is from 2 to 3 percent higher than when using $LC = 1$ for all choices of δ_R and δ_T.

Table I—Performance results for DTW factors

DTW Factors			Percent of Overall Recognition Accuracy	Average Distance Separation	Number of Occurrences of Maximum Distance Separation
LC	δ_R	δ_T			
0	0	0	76.4	0.044	356
0	4	0	76.4	0.043	355
0	0	0	76.9	0.045	483
0	4	4	77.2	0.044	445
1	0	0	74.4	0.039	291
1	4	0	73.7	0.039	294
1	0	4	74.8	0.040	377
1	4	4	74.2	0.039	343

(*ii*) The variation in recognition accuracy for a fixed value of LC (and varying δ_R and δ_T) is small.

(*iii*) The "best" recognition accuracy scores are for $LC = 0$, $\delta_R = 0$ or 4, and $\delta_T = 4$.

(*iv*) The average distance separations for $LC = 0$ are about 10 percent higher than for $LC = 1$.

(*v*) The best performances in terms of occurrences of maximum distance separation are for $LC = 0$, $\delta_T = 4$, and $\delta_R = 4$ and 0. Similarly for $LC = 1$, the cases $\delta_T = 4$ and $\delta_R = 0$ and 4 give the best performance in terms of this measurement.

3.2 Discussion

The results presented in Section 3.1 lead to the following two conclusions:

(*i*) The Itakura local constraints ($LC = 0$) yield better performance than the more general Type III local constraints for *all* choices of δ_R and δ_T, and for all three performance measurements.

(*ii*) The selection of $\delta_T = 4$ (with $\delta_R = 0$ or 4) is marginally better, in terms of recognition accuracy and average distance separation, than the selection of $\delta_T = 0$. However, in terms of maximum distance separation, the selection of $\delta_T = 4$ is significantly better than the selection of $\delta_T = 0$.

The first conclusion is expected in the sense that a great deal of earlier research has shown that any broadening of path regions invariably degrades performance for DTW algorithms,[4,6] since the improvement in score for the correct reference is generally more than offset by the improvement in score for the nearest incorrect reference. Although it was anticipated that the special construction of the reference patterns (i.e., from concatenated demisyllable units) might mitigate this general result, the data of Table I shows that this is not the case. It is especially notable that even for the case of two carefully chosen reference patterns in the recognizer, this general conclusion is still

valid. This result tends to lend some credence to the somewhat unusual testing methodology.

The second conclusion, although anticipated, is somewhat more difficult to explain. One expectation was that opening up the ending region (via nonzero values of δ_R and for δ_T) would lead to the same effect noted in conclusion one, namely, degraded performance. This was definitely not the case here. A second expectation was that if opening up the ending region was a good thing to do, *both* δ_R and δ_T would have to be nonzero. Although nonzero values of both δ_R and δ_T led to the best performance, it is seen that $\delta_R = 0$ was an almost equally good choice. The reason for this is the assymetry in the DTW implementation in which every single test frame is used in the warping path, but any given reference can be omitted entirely (since the path *can* jump by 2 frames). As such, at the end of the path, 2 of the 4 ending reference frames could be skipped with $\delta_R = 0$. Thus, the importance of a nonzero value of δ_R is greatly lessened, whereas a nonzero value of δ_T clearly leads to improved performance.

Additional analyses of the recognition results were performed to see if any simple correlations existed between cases in which improved recognitions occurred and specific linguistic events of the ends of the words, e.g., stops, nasals, fricatives, etc. No consistent and meaningful correlations were found. Hence, we conclude that the results of Table I are probably independent of the vocabulary items and the fact that the vocabulary was created from demisyllable tokens.

A final comment concerns the recognition accuracy achieved for the 1109-word vocabulary. Using whole word speaker-trained templates, Rabiner et al.[9] achieved a recognition accuracy of 79.2 percent across 6 talkers for the same 1109-word vocabulary. Thus, the best accuracy achieved using demisyllable-based templates (77.2 percent) compares favorably with the accuracy from whole word templates.

IV. SUMMARY

We have examined briefly the effects on isolated word recognition of three factors in the DTW alignment procedure. It was found that the Itakura local path constraints provided the best performance and that relaxing the ending constraint on the dynamic path also provided small, but consistent, improvements in performance.

REFERENCES

1. H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-26*, No. 6 (December 1978), pp. 575–82.
2. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-23*, No. 1 (February 1975), pp. 67–72.

3. G. M. White and R. B. Neely, "Speech Recognition Experiments with Linear Prediction, Bandpass Filtering, and Dynamic Programming," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-24,* No. 2 (April 1976), pp. 183–8.
4. L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson. "Considerations in Dynamic Time Warping for Discrete Word Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-26,* No. 6 (December 1978), pp. 575–82.
5. H. F. Silverman and N. R. Dixon, "State Constrained Dynamic Programming (SCDP) for Discrete Utterance Recognition," Proc. Int. Conf. ASSP (April 1980), pp. 169–72.
6. C. S. Myers, L. R. Rabiner, and A. E. Rosenberg, "Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-28,* No. 6 (December 1980), pp. 622–33.
7. A. E. Rosenberg, et al., "A Preliminary Study on the Use of Demisyllables in Automatic Speech Recognition," Proc. 1981 Int. Conf. Acoustics, Speech, and Signal Processing (March 1981), pp. 967–70.
8. C. K. Ogden, *Basic English: International Second Language,* New York: Harcourt, Brace, and World Inc., 1968.
9. L. R. Rabiner et al., unpublished work.